# TECHNICAL RESEARCH REPORT

Asymptotic optimality of the Round--Robin policy in multipath routing with resequencing

*by Konstantinos P. Tsoukatos, Armand M. Makowski*

**CSHCN TR 2005-2**
**(ISR TR 2005-79)**


*The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.*

**Web site http://www.isr.umd.edu/CSHCN/**

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **2005** | | **-** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Asymptotic optimality of the Round--Robin policy in multipath routing with resequencing** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Army Research Office,PO Box 12211,Research Triangle Park,NC,27709** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT
**see report**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | | **32** | |

# Asymptotic optimality of the Round–Robin policy in parallel queues with resequencing

KONSTANTINOS P. TSOUKATOS
Communications and Computer Engineering Department
University of Thessaly, Greece
ktsouk@inf.uth.gr

ARMAND M. MAKOWSKI *
Electrical Engineering Department and Institute for Systems Research,
University of Maryland, College Park, MD 20742
armand@isr.umd.edu

**Abstract**

We consider a model of a multipath routing system, where arriving customers are routed to a set of identical, parallel, single server queues, according to balancing policies operating without state information. After completion of service, customers are required to leave the system in their order of arrival, thus incurring an additional resequencing delay. We are interested in minimizing the end-to-end delay (including time at the resequencing buffer) experienced by arriving customers. To that end, we establish optimality of the Round–Robin routing assignment in two asymptotic regimes, namely heavy and light traffic: In heavy traffic, Round–Robin customer assignment is shown to achieve the smallest (in the increasing convex stochastic ordering) end-to-end delay amongst all routing policies operating without queue state information. In light traffic, and for the special case of Poisson arrivals, we show that Round–Robin is again an optimal (in the strong stochastic ordering) routing policy. We illustrate these and suggest other stochastic comparison results in a number of simulation examples.

*This work was prepared through collaborative participation in the Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U. S. Government.

# 1 Introduction

Many resource sharing systems act as disordering mechanisms in that their customers (e.g., packets in a communication network or tasks in a computer system) complete service in an order different from the order in which they entered the system. In some situations, a resequencing constraint is enforced by putting out-of-order customers into a *resequencing* buffer (after service completion) and delaying them until earlier customers catch up with them. The impact of such resequencing mechanisms on overall system performance is not always well understood. This can be attributed in part to the fact that even in the simplest of cases, the corresponding queueing models are not analytically tractable.

Earlier work on queueing systems with resequencing constraints is reported in [2, 5, 7, 8, 18]. Here we focus on a particular "disordering network" composed of $K$ identical servers operating in parallel, each attending to its own infinite capacity buffer and serving customers in FCFS order. Upon arrival to the system, customers are routed to one of the servers, with routing decisions being independent of the state of the system. Prominent among such routing mechanisms is the Round–Robin customer assignment. After service completion, each customer is required to leave the system in the order in which it arrived, thereby possibly experiencing an additional delay in the resequencing buffer. Of particular interest is the *end-to-end* delay experienced by customers; this quantity is defined as the sum of the time spent waiting in buffer for service (i.e., the waiting time), the time in service and the time spent in the resequencing buffer (i.e., the resequencing delay).

We assume customer arrivals to be modelled by a renewal process, and the service times to be i.i.d. and independent of the arrival process. We are interested not in evaluating various statistics of the customer end-to-end delay under these routing assignments, but in identifying the routing mechanism minimizing this quantity in some suitable sense. Intuitively, owing to its deterministic nature, the Round–Robin customer assignment should yield a smaller end-to-end delay than any other policy operating without state information, as yet another instance of the folk theorem that "determinism minimizes delays" [2] (and references therein). For instance, by the extremal property in [16, Prop. 6.3.1, p. 114], the steady state waiting time under Round–Robin is known to be smaller in the increasing convex ordering than that under the Bernoulli customer assignment. Thus, in line with this result, we would expect similar stochastic comparisons to hold in steady state for the end-to-end delay in the system of parallel queues with resequencing.

This paper is concerned with validating the optimality of the Round – Robin routing mechanism. For reasons to be explained shortly, we investigate the desired stochastic comparisons in two *asymptotic* regimes, namely, heavy and light loads. In particular, the heavy

traffic limit for the end-to-end delay under Round–Robin assignment is shown to be smaller, in the increasing convex stochastic ordering, than its counterpart under any other routing policy admitting a certain Functional Central Limit Theorem (in the form of Assumption (D) of Section 7). In light traffic, with Poisson arrivals, we show that Round–Robin routing achieves the smallest (in the strong stochastic ordering) end-to-end delay amongst all routing policies operating without queue state information (as understood in Assumption (A) of Section 3). These asymptotic comparisons are certainly compatible with the desired optimality across all traffic intensities, and give a strong indication of its validity; this is further reinforced by the limited simulation experiments reported in Section 12 for Poisson arrivals.

The difficulty in providing the desired comparison across the entire traffic range for stability lies in the fact that the end-to-end delay is a function of the entire workload *vector*. Under the enforced assumptions, the behavior of each queue in isolation is that of a $GI|GI|1$ queue under each of the routing mechanisms. This fact was exploited in the proof of Proposition 6.3.1 [16, p. 114] to compare the customer waiting times in the increasing convex stochastic ordering. However, in general the $K$ parallel single server queues are *not* independent, thereby precluding the derivation of closed-form expressions for the distributions of interest, even when such expressions are available for the corresponding $GI|GI|1$ queue.

This paper is organized as follows: The system model and basic performance metrics are presented in Section 2. Section 3 contains the basic statistical assumptions, and the existence of a steady state (or stationary) regime is discussed in Section 4. The heavy traffic problem is introduced in Section 5. This is followed in Section 6 by a summary of the heavy traffic methodology via weak convergence in function spaces. The assumptions for establishing the relevant heavy traffic limits are presented in Section 7, and the heavy traffic results are then developed in Section 8. These results are used to obtain stochastic comparisons in heavy traffic in Section 9. As a preamble to the light traffic results, we summarize the Reiman-Simon theory in Section 10. The light traffic calculations and subsequent comparisons are then discussed in Section 11. Finally, Section 12 discusses numerical examples that illustrate the results. A number of proofs are given in Appendix A.

A few words on the notation in use here: Throughout, let $K$ denote a given positive integer. Vectors are denoted in boldface and are always interpreted as *row* vectors. Moreover, the $k^{th}$ component of any element $\boldsymbol{x}$ in $\mathbb{R}^K$ is denoted by $x^k$, $k = 1, \ldots, K$, so that $\boldsymbol{x} \equiv (x^1, \ldots, x^K)$. A similar convention is used for random variables (rvs). For any scalar $x$ in $\mathbb{R}$, we denote by $\mathrm{mod}_K(x)$ the $\mathrm{mod}_K$-equivalent of $x$ in the interval $[0, K)$.

We use $\Longrightarrow_r$ to denote weak convergence (with $r$ going to infinity), and refer the reader to the monographs [3, 21] for additional information on weak convergence. We use $\leq_{st}$, $\leq_{cx}$ and $\leq_{icx}$ to denote inequality in the strong, convex and increasing convex stochastic orderings, respectively. Additional material on these orderings can be found in the monographs [14, 15, 16]. Equality in distribution between two rvs is denoted simply by $=_{st}$.

## 2 The system dynamics

To introduce the system of $K(\geq 2)$ parallel queues with resequencing, we start with the sequences of $\mathbb{R}_+$–valued rvs $\{\tau_{n+1},\ n = 0, 1, \ldots\}$ and $\{\sigma_n,\ n = 0, 1, \ldots\}$, and with the sequence of $\{1, \ldots, K\}$–valued rvs $\{\nu_n,\ n = 0, 1, \ldots\}$. With this last sequence we associate a new sequence of $\{0, 1\}^K$–valued rvs $\{\boldsymbol{u}_n,\ n = 0, 1, \ldots\}$ by setting

$$u_n^k = \delta(\nu_n, k), \quad k = 1, \ldots, K; \quad n = 0, 1, \ldots. \tag{1}$$

Throughout these quantities are given the following interpretation: For each $n = 0, 1, \ldots$, the interarrival time between the $n^{th}$ and the $(n+1)^{rst}$ customers is denoted by $\tau_{n+1}$ (with the convention that the $0^{th}$ customer arrives at time $t = 0$). The $n^{th}$ customer brings an amount of work that requires $\sigma_n$ units of execution time, while $\nu_n = k$ (or, equivalently, $u_n^k = 1$) indicates that the $n^{th}$ customer is routed to the $k^{th}$ queue.

We now define the performance measure of interest under the assumption that the system is empty at time $t = 0$. For each $k = 1, \ldots, K$, let $w_n^k$ represent the work remaining in the $k^{th}$ queue as seen by the $n^{th}$ customer just before entering the system. The $\mathbb{R}_+$–valued rvs $\{w_n^k,\ n = 0, 1, \ldots\}$ evolve according to the Lindley recursion

$$w_{n+1}^k = \left(w_n^k + u_n^k \sigma_n - \tau_{n+1}\right)^+, \quad n = 0, 1, \ldots. \tag{2}$$

with $w_0^k = 0$. Note that $w_n^k$ would be the waiting time of the $n^{th}$ customer, were she in fact to join the $k^{th}$ queue. Consequently, the time $d_n$ that the $n^{th}$ customer spends either in buffer waiting for service or in service is defined by

$$d_n := \sum_{k=1}^{K} u_n^k \left(w_n^k + \sigma_n\right) = \left(\sum_{k=1}^{K} u_n^k w_n^k\right) + \sigma_n, \quad n = 0, 1, \ldots. \tag{3}$$

As a global resequencing constraint is imposed on successive customers, a customer may spend time in a resequencing buffer after service completion, in order to wait for earlier customers which have been delayed. With this in mind we introduce the end-to-end delay of the $n^{th}$ customer as the rv $v_n$, which accounts for the time spent waiting in buffer

for service, the time in service and the time spent in the resequencing buffer. It is easy to see that

$$v_n = \max_{k=1,\ldots,K} \left( w_n^k + u_n^k \sigma_n \right), \quad n = 0, 1, \ldots . \tag{4}$$

## 3 Statistical assumptions

In this paper, we consider routing policies that balance the load among the $K$ queues by taking routing decisions which are independent of both the arrival process and the sequence of service times, but which are possibly dependent on past routing decisions. This is captured through the following set of assumptions:

**Assumption (A)** *(i) The three sequences* $\{\tau_{n+1},\ n = 0, 1, \ldots\}$, $\{\sigma_n,\ n = 0, 1, \ldots\}$ *and* $\{\nu_n,\ n = 0, 1, \ldots\}$ *are mutually independent; (ii) The* $\mathbb{R}_+$*–valued rvs* $\{\tau_{n+1},\ n = 0, 1, \ldots\}$ *form an i.i.d. sequence with finite variance* $\mathrm{var}\,[\tau]$*; (iii) The* $\mathbb{R}_+$*–valued rvs* $\{\sigma_n,\ n = 0, 1, \ldots\}$ *form an i.i.d. sequence with distribution* $G$ *and finite variance* $\mathrm{var}\,[\sigma]$*; (iv) The* $\{1, \ldots, K\}$*-valued rvs* $\{\nu_n,\ n = 0, 1, \ldots\}$ *form a stationary sequence with*

$$\mathbf{P}\left[\nu_n = k\right] = \frac{1}{K}, \quad k = 1, \ldots, K, \quad n = 0, 1, \ldots . \tag{5}$$

Under (A), let $\sigma$ and $\tau$ denote a pair of *independent* rvs; these are generic representatives of the i.i.d. sequences $\{\sigma_n,\ n = 0, 1, \ldots\}$ and $\{\tau_{n+1},\ n = 0, 1, \ldots\}$, respectively. Part (iv) of Assumption (A) guarantees that the $\{0, 1\}^K$-valued rvs $\{\boldsymbol{u}_n,\ n = 0, 1, \ldots\}$ given by (1) also form a stationary sequence. Any routing policy $\{\nu_n,\ n = 0, 1, \ldots\}$ satisfying Assumption (A) is said to be *admissible*.

Of central interest is the Round–Robin customer assignment. This is implemented through the routing rvs $\{\nu_n^R,\ n = 0, 1, \ldots\}$ given by

$$\nu_n^R = \mathrm{mod}_K(\nu^\star + n - 1) + 1, \quad n = 0, 1, \ldots \tag{6}$$

with rv $\nu^\star$ uniform over $\{1, \ldots, K\}$, i.e.,

$$\mathbf{P}\left[\nu^\star = k\right] = \frac{1}{K}, \quad k = 1, \ldots, K. \tag{7}$$

With this definition, the $0^{th}$ customer is randomly routed to one of the $K$ queues. We also specialize several results to the Bernoulli customer assignment, which is characterized by the routing rvs $\{\nu_n^B,\ n = 0, 1, \ldots\}$ being i.i.d. rvs with common distribution given by

$$\mathbf{P}\left[\nu_n^B = k\right] = \frac{1}{K}, \quad k = 1, \ldots, K, \quad n = 0, 1, \ldots \tag{8}$$

Under either assignment, the routing rvs $\{\nu_n,\ n = 0, 1, \ldots\}$ form a stationary sequence of $\{1, \ldots, K\}$-valued rvs with common marginal distribution given by (5), hence both Round–Robin and Bernoulli are admissible routing policies. Following the usage started in (8) and (6), quantities associated with the system under the Bernoulli and Round–Robin customer assignments are distinguished by the superscript $B$ and $R$, respectively. Omission of the superscript will reflect the fact that the discussion holds for any admissible policy.

## 4 Steady state regimes and stochastic comparisons

Under any admissible routing policy, the system is stable if and only if each of the $K$ queues is stable. It is a simple matter to check that this will happen if and only the *(server) utilization* $\rho$ is less than one, namely

$$\rho := \frac{\mathbf{E}[\sigma]}{K\mathbf{E}[\tau]} < 1. \tag{9}$$

The precise formulation of this stability property is contained in a multi-dimensional analog of Loynes' result [10] given below; its proof is straightforward and omitted for the sake of brevity.

In order to state the result we need to expand the setup of Assumption (A) to *bi-infinite* sequences, possibly by enlarging the underlying probability space in the usual manner. Thus, whenever Assumption (A) holds, we are to understand the following: (i) There exist three bi-infinite sequences $\{\tau_{n+1},\ n = 0, \pm 1, \pm 2, \ldots\}$, $\{\sigma_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ and $\{\nu_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ which are mutually independent; (ii) The $\mathbb{R}_+$–valued rvs $\{\tau_{n+1},\ n = 0, \pm 1, \pm 2, \ldots\}$ form an i.i.d. sequence with finite variance $\mathrm{var}[\tau]$; (iii) The $\mathbb{R}_+$–valued rvs $\{\sigma_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ form an i.i.d. sequence with distribution $G$ and finite variance $\mathrm{var}[\sigma]$; (iv) The $\{1, \ldots, K\}$-valued rvs $\{\nu_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ form a stationary sequence with

$$\mathbf{P}[\nu_n = k] = \frac{1}{K}, \quad k = 1, \ldots, K, \quad n = 0, \pm 1, \pm 2, \ldots. \tag{10}$$

Fix $k = 1, 2, \ldots, K$. We define the sequence of partial sums $\{s_n^k,\ n = 0, \pm 1, \pm 2, \ldots\}$ by

$$s_0^k := 0; \quad s_n^k := \begin{cases} \displaystyle\sum_{m=1}^{n} \xi_m^k & \text{if } n = 1, 2, \ldots \\ \\ \displaystyle\sum_{m=n+1}^{0} \xi_m^k & \text{if } n = -1, -2, \ldots \end{cases} \tag{11}$$

where for each $n = 0, \pm 1, \pm 2, \ldots$, we have set

$$\xi_{n+1}^k := u_n^k \sigma_n - \tau_{n+1} \tag{12}$$

with

$$u_n^k = \delta(\nu_n, k). \tag{13}$$

**Proposition 4.1** *Under Assumption (A), the stability condition (9) ensures that* $\boldsymbol{w}_n \Longrightarrow_n \boldsymbol{w}_\infty = (w_\infty^1, \ldots, w_\infty^K)$ *and* $v_n \Longrightarrow_n v_\infty$ *with*

$$w_\infty^k := \sup\left(s_m^k,\ m = 0, -1, \ldots\right), \quad k = 1, \ldots, K \tag{14}$$

*and*

$$v_\infty := \max_{k=1,\ldots,K}\left(w_\infty^k + u_0^k \sigma_0\right). \tag{15}$$

We refer to $\boldsymbol{w}_\infty$ and $v_\infty$ as the stationary workload vector and end-to-end delay, respectively. To shed light into their representations (14) and (15), we note the monotone convergence

$$w_\infty^k = \lim_{n\to\infty} \widetilde{w}_n^k, \quad k = 1, \ldots, K \tag{16}$$

where we have set

$$\widetilde{w}_n^k := \max\left(s_m^k,\ m = 0, -1, \ldots, -n\right), \quad n = 1, 2, \ldots \tag{17}$$

For each $n = 1, 2, \ldots$, stationarity implies that

$$w_n^k =_{st} \widetilde{w}_n^k \quad \text{and} \quad v_n =_{st} \widetilde{v}_n \tag{18}$$

with

$$\widetilde{v}_n := \max_{k=1,\ldots,K}\left(\widetilde{w}_n^k + u_0^k \sigma_0\right). \tag{19}$$

This identification reflects the usual Loynes's "backward in time" arguments to construct the stationary regime [10].

Under the enforced assumptions, it holds that $w_\infty^1 =_{st} \ldots =_{st} w_\infty^K$, and we readily conclude from (3) and (14) that

$$d_n \Longrightarrow_n d_\infty \quad \text{with} \quad d_\infty := \left(\sum_{k=1}^K u_0^k w_\infty^k\right) + \sigma_0. \tag{20}$$

It is well known [16, Prop. 6.3.1, p. 114] that $w_\infty^{1,R} \leq_{icx} w_\infty^{1,B}$, so that $d_\infty^R \leq_{icx} d_\infty^B$ in view of (20). It is therefore natural to wonder whether such a comparison also holds between the end-to-end delays $v_\infty^R$ and $v_\infty^B$, namely $v_\infty^R \leq_{icx} v_\infty^B$. It turns out that an answer to this question is much more elusive, and prompts us instead to seek such stochastic comparisons in the limiting regimes of heavy and light traffic.

## 5 The heavy traffic problem

We seek a characterization of the system of parallel queues with resequencing in the case where it is almost fully utilized, i.e., the utilization, though less than one, is very close to unity. This heavy traffic characterization entails model simplifications which allow for subsequent comparison between of the end-to-end delays under Round–Robin and any other admissible routing policy.

To do so, we embed the system of parallel queues into a parametric family of like queueing systems, indexed by an integer (say $r$), with the property that the utilization $\rho_r$ of the $r^{th}$ system tends to the critical value 1 as $r$ goes to infinity. These systems differ only through their arrival sequences. Specifically, for each $r = 1, 2, \ldots$, the $r^{th}$ system of parallel queues with resequencing is one driven by the sequence of service times $\{\sigma_n,\ n = 0, \pm 1, \pm 2, \ldots\}$, the sequence of routing rvs $\{\nu_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ and the arrival sequence $\{\tau^r_{n+1},\ n = 0, \pm 1, \pm 2, \ldots\}$ under Assumption (A). Quantities associated with the $r^{th}$ system are superscripted by $r$.

We take $\rho_r < 1$, or equivalently, $\mathbf{E}[\sigma] < K\mathbf{E}[\tau^r]$, to ensure stability of the $r^{th}$ system, so that $\boldsymbol{w}^r_n \Longrightarrow_n \boldsymbol{w}^r_\infty$ and $v^r_n \Longrightarrow_n v^r_\infty$ by Proposition 4.1. However, we drive the family of systems to *heavy traffic* by assuming that

$$\lim_{r\to\infty} \mathbf{E}[\tau^r] = \frac{1}{K}\,\mathbf{E}[\sigma]\,. \tag{21}$$

Clearly, as $r$ goes to infinity, the components of the stationary workload vector $\boldsymbol{w}^r_\infty$ grow unbounded. It is therefore appropriate to seek a scaling sequence $\{\alpha_r,\ r = 1, 2, \ldots\}$ with $\lim_{r\to\infty}\alpha_r = \infty$ such that the convergence in distribution

$$\alpha_r{}^{-1}\boldsymbol{w}^r_\infty \Longrightarrow_r \boldsymbol{w}_{\rm HT} \tag{22}$$

takes place to some $\mathbb{R}^K_+$-valued rv $\boldsymbol{w}_{\rm HT} = (w^1_{\rm HT}, \ldots, w^K_{\rm HT})$. It then follows from (15) that

$$\alpha_r{}^{-1}v^r_\infty \Longrightarrow_r v_{\rm HT} \tag{23}$$

with $\mathbb{R}_+$-valued rv $v_{\rm HT}$ determined through the relation

$$v_{\rm HT} = \max_{k=1,\ldots,K} w^k_{\rm HT}. \tag{24}$$

It is customary to refer to $\boldsymbol{w}_{\rm HT}$ and $v_{\rm HT}$ as the heavy traffic limits.

## 6 The heavy traffic methodology

In order to identify the scaling sequence and the heavy traffic limits in (22)-(24), we take the indirect approach based on diffusion limits whereby the quantities of interest are rescaled both in the time and state space variables [1, 19].

To introduce the basic ideas, fix $r = 1, 2, \ldots$ and $k = 1, \ldots, K$. The identification (18) leads in the usual manner [19, 21] to considering the rvs

$$\begin{aligned} \widetilde{w}^{r,k}_{[rt]} &= \max\left(s^{r,k}_m,\ m = 0, -1, \ldots, -[rt]\right) \\ &= \max\left(s^{r,k}_{-[ru]},\ 0 \le u \le t\right), \quad t \ge 0 \end{aligned} \tag{25}$$

and

$$\widetilde{v}^{r}_{[rt]} = \max_{k=1,\ldots,K}\left(\widetilde{w}^{r,k}_{[rt]} + u^k_0\sigma_0\right), \quad t \ge 0. \tag{26}$$

Next, we define the $\mathbb{R}$–valued processes $\{S^{r,k}(t),\ t \ge 0\}$ and $\{\widetilde{W}^{r,k}(t),\ t \ge 0\}$ by

$$S^{r,k}(t) := \frac{1}{\sqrt{r}}\left(s^{r,k}_{-[rt]} - \mathbf{E}\left[s^{r,k}_{-[rt]}\right]\right) \quad \text{and} \quad \widetilde{W}^{r,k}(t) := \frac{\widetilde{w}^{r,k}_{[rt]}}{\sqrt{r}}, \quad t \ge 0. \tag{27}$$

The $\mathbb{R}^K$-valued processes with components given by (27) are denoted by $\{\boldsymbol{S}(t),\ t \ge 0\}$ and $\{\widetilde{\boldsymbol{W}}^r(t),\ t \ge 0\}$, respectively.

With the help of (25) we note that

$$\widetilde{W}^{r,k}(t) = \sup_{0\le u\le t}\left(S^{r,k}(u) - \gamma^{r,k}(u)\right), \quad t \ge 0 \tag{28}$$

where we have defined the function $\gamma^{r,k} : \mathbb{R}_+ \to \mathbb{R}$ by

$$\gamma^{r,k}(t) := -\frac{1}{\sqrt{r}}\,\mathbf{E}\left[s^{r,k}_{-[rt]}\right], \quad t \ge 0. \tag{29}$$

Finally, we define the $\mathbb{R}$–valued process $\{\widetilde{V}^r(t),\ t \ge 0\}$ by

$$\widetilde{V}^r(t) := \max_{k=1,\ldots,K}\left(\widetilde{W}^{r,k}(t) + u^k_0\frac{\sigma_0}{\sqrt{r}}\right), \quad t \ge 0. \tag{30}$$

The next step consists in letting $r$ go to infinity in these definitions, with limits understood in the sense of weak convergence on function spaces [3]: For each $T > 0$, let $D[0,T]^K$ denote the space of mappings $[0,T] \to \mathbb{R}^K$ which are right continuous with left limits; as usual the vector space $D[0,T]^K$ is equipped with the Skorokhod topology. Consider the sequence of $\mathbb{R}^K$–valued processes $\{\boldsymbol{X}^r(t),\ t \ge 0\}$ $(r = 1, 2, \ldots)$ with sample paths in $D[0,\infty)^K$. Whenever, for each $T > 0$, the weak convergence

$$\{\boldsymbol{X}^r(t),\ 0 \le t \le T\} \Longrightarrow_r \{\boldsymbol{X}(t),\ 0 \le t \le T\} \quad \text{in } D[0,T]^K$$

takes place (under the Skorokhod topology) for some $\mathbb{R}^K$-valued process $\{\boldsymbol{X}(t),\ t \geq 0\}$ with sample paths in $D[0,\infty]^K$, we simply write

$$\{\boldsymbol{X}^r(t),\ t \geq 0\} \Longrightarrow_r \{\boldsymbol{X}(t),\ t \geq 0\}.$$

It is plain from (28) and (30) that the processes $\{\widetilde{\boldsymbol{W}}^r(t),\ t \geq 0\}$ and $\{\widetilde{V}^r(t),\ t \geq 0\}$ are obtained through non-linear functionals on the process $\{\boldsymbol{S}^r(t),\ t \geq 0\}$. This observation suggests that the limit properties of the latter process will determine those of the former group. To emphasize this point further, for each $T > 0$ we introduce the *supremum mapping* $M_T : D[0,T] \to D[0,T]$ defined by

$$M_T(x)(t) := \sup_{0 \leq u \leq t} x(u), \quad 0 \leq t \leq T \tag{31}$$

at element $x$ of $D[0,T]$. In particular, relation (28) now becomes

$$\widetilde{W}^{r,k}(t) = M_T(\{S^{r,k}(u) - \gamma^{r,k}(u),\ 0 \leq u \leq T\})(t), \quad 0 \leq t \leq T. \tag{32}$$

It is well known that the supremum mapping $M_T : D[0,T] \to D[0,T]$ is continuous under the Skorokhod topology [20, Thm 6.4, p. 81].

# 7 The heavy traffic assumptions

A number of additional assumptions are needed to carry out the discussion. Assumption (B) below complements (21); it is enforced thereafter and ensures that the system described by (28)–(30) is driven to heavy traffic at the appropriate speed.

**Assumption (B)** *The sequence of generic interarrival times* $\{\tau^r,\ r = 1, 2, \ldots\}$ *satisfies*

$$\lim_{r \to \infty} \sqrt{r} \left( \frac{1}{K}\, \mathbf{E}\left[\sigma\right] - \mathbf{E}\left[\tau^r\right] \right) = -\gamma$$

*for some* $\gamma > 0$.

Under Assumption (A), for each $k = 1, \ldots, K$, we have

$$\mathbf{E}\left[s^{r,k}_{[rt]}\right] = \mathbf{E}\left[\sigma\right] \sum_{n=0}^{[rt]-1} \mathbf{E}\left[u^k_n\right] - [rt]\mathbf{E}\left[\tau^r\right], \quad t \geq 0.$$

Hence, Assumption (B) readily leads via (29) to

$$\lim_{r \to \infty} \gamma^{r,k}(t) = \gamma t, \quad t \geq 0 \tag{33}$$

under any admissible routing policy.

We follow up with a technical assumption on the arrival sequence:

**Assumption (C)** *The rvs* $\{|\tau^r|^2,\ r = 1, 2, \ldots\}$ *are uniformly integrable rvs with* $\zeta^2 := \lim_{r\to\infty} \mathrm{var}\,[\tau^r] > 0$.

Assumption (C) ensures the validity of a version of Donsker's Theorem [3, Thm. 16.1, p. 137] in the form of the following Functional Central Limit Theorem (FCLT), namely

$$\left\{ \frac{1}{\sqrt{r}} \sum_{n=0}^{[rt]-1} (\tau_{n+1}^r - \mathbf{E}\,[\tau^r]),\ t \geq 0 \right\} \Longrightarrow_r \{\zeta B(t), t \geq 0\}. \tag{34}$$

with $\{B(t),\ t \geq 0\}$ a standard one-dimensional Brownian motion. The uniform integrability in Assumption (C) is used to validate the appropriate Lindeberg's condition [3, Eqn. (7.3), p. 42].

The heavy traffic results will be established only for those admissible routing policies which admit a FCLT in a sense which we now describe: Here, and throughout, let $\mathbf{1}$ denote the element $(1, \ldots, 1)$ in $\mathbb{R}^K$. Next, consider an admissible routing policy $\{\nu_n,\ n = 0, 1, \ldots\}$ with corresponding sequence $\{\boldsymbol{u}_n,\ n = 0, 1, \ldots\}$. For each $r = 1, 2, \ldots$, we define the $\mathbb{R}^K$-valued process $\{\boldsymbol{U}^r(t),\ t \geq 0\}$ by

$$\boldsymbol{U}^r(t) := \frac{1}{\sqrt{r}} \sum_{n=0}^{[rt]-1} \left( \boldsymbol{u}_n \sigma_n - \frac{\mathbf{E}\,[\sigma]}{K} \mathbf{1} \right),\ t \geq 0. \tag{35}$$

**Assumption (D)** *Consider an admissible routing policy* $\{\nu_n,\ n = 0, 1, \ldots\}$. *The corresponding sequence* $\{\boldsymbol{u}_n,\ n = 0, 1, \ldots\}$ *satisfies the FCLT in the form*

$$\{\boldsymbol{U}^r(t),\ t \geq 0\} \Longrightarrow_r \{\Gamma^{1/2} \boldsymbol{A}(t),\ t \geq 0\} \tag{36}$$

*where* $\{\boldsymbol{A}(t),\ t \geq 0\}$ *is a* $K$*–dimensional standard Brownian motion, and* $\Gamma$ *is the* $K \times K$ *covariance matrix determined by*

$$\Gamma_{k\ell} = \lim_{r\to\infty} \frac{1}{r} \mathrm{cov} \left[ \sum_{n=0}^{r-1} u_n^k \sigma_n, \sum_{m=0}^{r-1} u_m^\ell \sigma_m \right], \quad k, \ell = 1, \ldots, K. \tag{37}$$

An admissible routing policy is said to be *HT–admissible* if Assumption (D) holds. The role of the matrix $\Gamma$ in (36) can be made more transparent as follows: For each $t > 0$, (37) is equivalent to

$$\mathrm{var} \left[ \left( \sum_{n=0}^{[rt]-1} \boldsymbol{u}_n \sigma_n \right) \boldsymbol{\theta}^T \right] \sim rt \cdot \boldsymbol{\theta} \Gamma \boldsymbol{\theta}^T, \quad \boldsymbol{\theta} \in \mathbb{R}^K \tag{38}$$

for $r$ becoming large. Therefore, with the Cramér-Wold device [3, p. 48] in mind, we now see that (36) implies the convergence

$$\frac{\left(\sum_{n=0}^{[rt]-1}\left(\boldsymbol{u}_n\sigma_n - \frac{\mathbf{E}[\sigma]}{K}\mathbf{1}\right)\right)\boldsymbol{\theta}^T}{\sqrt{\mathrm{var}\left[\left(\sum_{n=0}^{[rt]-1}\boldsymbol{u}_n\sigma_n\right)\boldsymbol{\theta}^T\right]}} \Longrightarrow_r U, \quad \boldsymbol{\theta}\in\mathbb{R}^K \tag{39}$$

where $U$ denotes the standard zero-mean unit variance Gaussian rv. This last convergence is in the form of a Central Limit Theorem for identically distributed rvs which are not necessarily independent; this is an extensively studied problem for which results abound in the literature. As HT–admissibility represents the functional form of such CLT results, it is expected to be satisfied by a large class of admissible routing policies. The reader is refered to Section 4.4 of the monograph by Whitt [21] for an overview of some of the possibilities.

The next result determines the impact of the routing policy on the covariance matrix $\Gamma$. Its proof is available in Appendix A.1.

**Lemma 7.1** *The $K\times K$ covariance matrix $\Gamma$ given by (37) exists if and only if the $K\times K$ covariance matrix $\widetilde{\Gamma}$ determined by*

$$\widetilde{\Gamma}_{k\ell} = \lim_{r\to\infty}\frac{1}{r}\mathrm{cov}\left[\sum_{n=0}^{r-1}u_n^k, \sum_{m=0}^{r-1}u_m^\ell\right], \quad k,\ell = 1,\ldots,K \tag{40}$$

*exists, in which case*

$$\Gamma_{k\ell} = \widetilde{\Gamma}_{k\ell}\cdot\mathbf{E}\left[\sigma\right]^2 + \delta(k,\ell)\frac{\mathrm{var}\left[\sigma\right]}{K}, \quad k,\ell = 1,\ldots,K. \tag{41}$$

# 8 The heavy traffic limits

We begin by discussing the HT–admissibility of Round–Robin and Bernoulli routing. Proofs are available in Appendix A.2.

**Proposition 8.1** *Both Bernoulli and Round–Robin routing policies satisfy Assumption (D) with covariance matrices (40) given by*

$$\widetilde{\Gamma}^B_{k\ell} = \delta(k,\ell)\frac{1}{K}\left(1-\frac{1}{K}\right), \quad k,\ell = 1,\ldots,K, \tag{42}$$

*and*

$$\widetilde{\Gamma}^R_{k\ell} = 0, \quad k,\ell = 1,\ldots,K, \tag{43}$$

*respectively.*

We are now ready to develop the requisite heavy traffic limits under an arbitrary HT–admissible routing policy.

**Proposition 8.2** *Consider an HT-admissible routing policy* $\{\nu_n,\ n = 0, 1, \ldots\}$ *under the Assumptions (A)–(D). The convergence*

$$\{\boldsymbol{S}^r(t),\ t \geq 0\} \Longrightarrow_r \{\boldsymbol{S}(t),\ t \geq 0\} \tag{44}$$

*holds with*

$$\boldsymbol{S}(t) := \boldsymbol{B}(t)\Sigma^{1/2},\ t \geq 0 \tag{45}$$

*where* $\{\boldsymbol{B}(t),\ t \geq 0\}$ *denotes a* $K$*–dimensional standard Brownian motion and the* $K \times K$ *covariance matrix* $\Sigma$ *is given by*

$$\Sigma_{k\ell} = \Gamma_{k\ell} + \zeta^2, \quad k, \ell = 1, \ldots, K. \tag{46}$$

With the help of (41), we can rewrite (46) more compactly in matrix form as

$$\Sigma = \Gamma + \zeta^2 E = \widetilde{\Gamma} \cdot \mathbf{E}\left[\sigma\right]^2 + \frac{\mathrm{var}\left[\sigma\right]}{K} I + \zeta^2 E \tag{47}$$

where $E$ is the $K \times K$ covariance matrix $\mathbf{1}^T\mathbf{1}$, and $I$ denotes the identity matrix on $\mathbb{R}^K$.

**Proof.** Under Assumption (C), (34) can be given a multi-dimensional version in the form

$$\left\{ \frac{1}{\sqrt{r}} \sum_{n=0}^{[rt]-1} (\tau^r_{n+1} - \mathbf{E}\left[\tau^r\right])\mathbf{1},\ t \geq 0 \right\} \Longrightarrow_r \{\zeta B(t)\mathbf{1}, t \geq 0\}. \tag{48}$$

Given the enforced independence assumptions, the Brownian motions $\{B(t),\ t \geq 0\}$ above and $\{\boldsymbol{A}(t),\ t \geq 0\}$ of (36) can be taken to be mutually independent. Hence, for each $t > 0$, we have[1]

$$\begin{aligned} \mathrm{cov}\left[\boldsymbol{A}(t)\Gamma^{1/2} + \zeta B(t)\mathbf{1}\right] &= \mathrm{cov}\left[\boldsymbol{A}(t)\Gamma^{1/2}\right] + \zeta^2 \mathrm{cov}\left[B(t)\mathbf{1}\right] \\ &= \Gamma^{1/2}\mathrm{cov}\left[\boldsymbol{A}(t)\right]\Gamma^{1/2} + \zeta^2 \mathbf{1}^T \mathrm{var}\left[B(t)\right]\mathbf{1} \\ &= \left(\Gamma + \zeta^2 \mathbf{1}^T\mathbf{1}\right) t \end{aligned} \tag{49}$$

since $\mathrm{var}\left[B(t)\right] = t$ and $\mathrm{cov}\left[\boldsymbol{A}(t)\right] = I \cdot t$. In short, $\mathrm{cov}\left[\boldsymbol{A}(t)\Gamma^{1/2} + \zeta B(t)\mathbf{1}\right] = \Sigma \cdot t$, and the identification

$$\{\boldsymbol{A}(t)\Gamma^{1/2} + \zeta B(t)\mathbf{1},\ t \geq 0\} =_{st} \{\boldsymbol{B}(t)\Sigma^{1/2},\ t \geq 0\} \tag{50}$$

[1]Recall that $\Gamma$ and $\Gamma^{1/2}$ are symmetric matrices.

follows by the usual characterization of Brownian motion.

To conclude, recall the definition (27) and (35) (together with (11) and (12)). Making use of (36), (33) and (48), we find that

$$\left\{ \boldsymbol{U}^r(t) + \frac{1}{\sqrt{r}} \sum_{n=0}^{[rt]-1} (\tau_{n+1}^r - \mathbf{E}\,[\tau^r])\mathbf{1},\ t \geq 0 \right\} \Longrightarrow_r \{\boldsymbol{B}(t)\Sigma^{1/2},\ t \geq 0\}.$$

This an easy consequence of the Continuous Mapping Theorem [3, p. 29] [19, p. 320] when coupled with Theorem 3.2 in [3, p. 21]. This completes the proof of (44) as we note the stochastic equivalencce

$$\{\boldsymbol{S}^r(t),\ t \geq 0\} =_{st} \left\{ \boldsymbol{U}^r(t) + \frac{1}{\sqrt{r}} \sum_{n=0}^{[rt]-1} (\tau_{n+1}^r - \mathbf{E}\,[\tau^r])\mathbf{1},\ t \geq 0 \right\}$$

which is easily validated under the enforced assumptions (A). ■

Whenever convergence (44) holds, we conclude that

$$\{\widetilde{\boldsymbol{W}}^r(t),\ t \geq 0\} \Longrightarrow_r \{\widetilde{\boldsymbol{W}}(t),\ t \geq 0\} \tag{51}$$

with

$$\widetilde{W}^k(t) := \sup_{0 \leq u \leq t} \left( S^k(u) - \gamma u \right), \quad k = 1, \ldots, K,\ t \geq 0. \tag{52}$$

This convergence is a direct consequence of the Continuous Mapping Theorem [3, p. 29] [19, p. 320] (via (32) and (33)) given the aforementioned continuity of the supremum mapping (31) in the Skorokhod topology; details for the multi-dimensional extension are standard and left to the interested reader.

Now, going back to (30) we see that

$$0 \leq \widetilde{V}^r(t) - \max_{k=1,\ldots,K} \widetilde{W}^{r,k}(t) \leq \frac{\sigma_0}{\sqrt{r}}, \quad t \geq 0 \tag{53}$$

for each $r = 1, 2, \ldots$. Applying the Continuous Mapping Theorem [3, p. 29], in conjunction with (51), yields $\{\max_{k=1,\ldots,K} \widetilde{W}^{r,k}(t),\ t \geq 0\} \Longrightarrow_r \{\widetilde{V}(t),\ t \geq 0\}$ where

$$\widetilde{V}(t) := \max_{k=1,\ldots,K} \widetilde{W}^k(t), \quad t \geq 0. \tag{54}$$

The conclusion

$$\{\widetilde{V}^r(t),\ t \geq 0\} \Longrightarrow_r \{\widetilde{V}(t),\ t \geq 0\}$$

is now a straightforward consequence of the inequality (53) and of the Convergence Together Theorem [3, Thm. 4.1, p. 25].

Let $t$ go to infinity in (52) and (54). Standard monotonicity arguments yield the stationary heavy traffic workload vector $\widetilde{\boldsymbol{W}}$ and end-to-end delay $\widetilde{V}$ as the limiting rvs $\lim_{t\to\infty} \widetilde{\boldsymbol{W}}(t) = \widetilde{\boldsymbol{W}}$ and $\lim_{t\to\infty} \widetilde{V}(t) = \widetilde{V}$ with

$$\widetilde{W}^k := \sup_{t\geq 0} \Big( S^k(t) - \gamma t \Big), \quad k = 1, \ldots, K \tag{55}$$

and

$$\widetilde{V} := \max_{k=1,\ldots,K} \widetilde{W}^k. \tag{56}$$

The finiteness of these rvs is discussed in the course of proving Proposition 9.3.

Going back to (22)-(24), we see that we can select $\alpha_r = \sqrt{r}$, and the arguments in the one-dimensional case [1] readily extend to yield the identification

$$\widetilde{\boldsymbol{W}} =_{st} \boldsymbol{w}_{\rm HT} \quad \text{and} \quad \widetilde{V} =_{st} v_{\rm HT}, \tag{57}$$

thereby validating the use of diffusion limits to secure heavy traffic results.

# 9 Heavy traffic optimality of Round – Robin routing

We are interested in a stochastic comparison in heavy traffic between the end-to-end delay under Round–Robin and that of any other HT–admissible routing policy. Our first step in this direction is a characterization of the convex ordering for Gaussian rvs due to Müller [11, Thm. 3.3].

**Proposition 9.1** *Let $\boldsymbol{X}$ and $\boldsymbol{X}'$ denote two normally distributed $\mathbb{R}^d$-valued rvs, say $\boldsymbol{X} =_{st} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{X}' =_{st} \mathcal{N}(\boldsymbol{\mu}', \Sigma')$, respectively. Then $\boldsymbol{X} \leq_{cx} \boldsymbol{X}'$ if and only if $\boldsymbol{\mu} = \boldsymbol{\mu}'$ and the $d \times d$ matrix $\Sigma' - \Sigma$ is positive semi-definite.*

Consider now an HT-admissible routing policy $\{\nu_n,\ n = 0, 1, \ldots\}$ under the Assumptions (A)–(D). To apply Proposition 9.1 we fix some $n = 1, 2, \ldots$, and with any ordered $n$-tuple $t_1 < \ldots < t_n$ in $\mathbb{R}_+$, we associate the $nK$–dimensional rv $\boldsymbol{S}(t_1, \ldots, t_n)$ given by

$$\boldsymbol{S}(t_1, \ldots, t_n) := (\boldsymbol{S}(t_1), \ldots, \boldsymbol{S}(t_n))$$

where $\{\boldsymbol{S}(t),\ t \geq 0\}$ is the $\mathbb{R}^K$-valued limiting process (45) identified in Proposition 8.2. The comparison result that follows is established in Appendix A.3. As usual, the superscript $R$ indicates that the corresponding quantity is evaluated under the Round–Robin policy.

**Proposition 9.2** *Consider an HT-admissible routing policy under the Assumptions (A)–(D). For all $n = 1, 2, \ldots$, and $0 \leq t_1 < \ldots < t_n$, it holds that*

$$\boldsymbol{S}^R(t_1, \ldots, t_n) \leq_{cx} \boldsymbol{S}(t_1, \ldots, t_n). \tag{58}$$

Comparison (58) leads to the desired stochastic comparison between the stationary heavy traffic end-to-end delays under Round–Robin and any other HT–admissible routing policy.

**Proposition 9.3** *For any HT-admissible routing policy under Assumptions (A)–(D), it holds that*

$$\widetilde{\boldsymbol{W}}^R \leq_{icx} \widetilde{\boldsymbol{W}} \quad \text{and} \quad \widetilde{V}^R \leq_{icx} \widetilde{V}. \tag{59}$$

In particular, making use of the identification (57), we obtain the heavy traffic comparison $v_{\mathrm{HT}}^R \leq_{icx} v_{\mathrm{HT}}$, which suggests the validity of the comparison $v_\infty^{r,R} \leq_{icx} v_\infty^r$ for sufficiently large $r$.

**Proof.** Fix $n = 1, 2, \ldots$. In the notation of Proposition 9.2, for any ordered $n$-tuple $0 \leq t_1 < \ldots < t_n$ (with the convention $t_0 = 0$), let $\boldsymbol{M}(t_1, \ldots, t_n)$ denote the $\mathbb{R}^K$-valued rv with components

$$M^k(t_1, \ldots, t_n) := \max\Big(S^k(t_i) - \gamma t_i,\ i = 0, 1, \ldots, n\Big), \quad k = 1, \ldots, K.$$

We conclude from (58) that

$$\boldsymbol{M}^R(t_1, \ldots, t_n) \leq_{icx} \boldsymbol{M}(t_1, \ldots, t_n) \tag{60}$$

owing to the fact that comparisons in the convex increasing ordering are preserved under increasing convex mappings [15].

Before we can make use of these comparisons, we need some preparatory work: Fix $k = 1, \ldots, K$ and note from the definitions (52) and (55) that

$$0 \leq \widetilde{W}^k(t) \leq \widetilde{W}^k, \quad t \geq 0. \tag{61}$$

The process $\{S^k(t), \geq t \geq 0\}$ is statistically indistinguishable from $\{\sigma_k B^k(t), \geq t \geq 0\}$ where $\sigma_k = \sqrt{\Sigma_{kk}}$, and it is a simple matter to check from (47) that $\sigma_k > 0$. Thus, by standard results on the supremum of Brownian motion [1, 3, 19], the rv $\widetilde{W}^k$ is exponentially

distributed, and the rvs $\{\widetilde{W}^k(t),\ t \geq 0\}$ are therefore uniformly integrable by virtue of (61). The uniform integrability of the $\mathbb{R}^K$-valued rvs $\{\widetilde{\boldsymbol{W}}(t),\ t \geq 0\}$ follows with $\lim_{t\to\infty} \widetilde{\boldsymbol{W}}(t) = \widetilde{\boldsymbol{W}}$.

Next, fix $t > 0$. For each $p = 1, 2, \ldots$, take $t_{p,j} = j2^{-p}t$ with $j = 0, \ldots, 2^p$, and write

$$M^k(t;p) := M^k(t_{p,0}, t_{p,1}, \ldots, t_{p,2^p}).$$

In this notation, comparison (60) yields

$$\boldsymbol{M}^R(t;p) \leq_{icx} \boldsymbol{M}(t;p), \quad p = 1, 2, \ldots. \tag{62}$$

Note that $\boldsymbol{M}(t;p) \leq \boldsymbol{M}(t;p+1) \leq \widetilde{\boldsymbol{W}}(t)$ componentwise for all $p = 0, 1, \ldots$ and that for each $k = 1, \ldots, K$, we have $\lim_{p\to\infty} M^k(t;p) = \widetilde{W}^k(t)$ monotonically from below by a simple continuity argument. Letting $p$ go to infinity in (62) yields

$$\widetilde{\boldsymbol{W}}^R(t) \leq_{icx} \widetilde{\boldsymbol{W}}(t) \tag{63}$$

by the uniform integrability of the rvs $\{\widetilde{\boldsymbol{W}}(t),\ t \geq 0\}$ [16, Prop. 1.3.2, p. 10]. As we let $t$ go to infinity in (63), we get the first convergence in (59) [16, Prop. 1.3.2, p. 10] upon invoking again the aforementioned uniform integrability.

The second convergence is now immediate from the fact the convex increasing mapping $\mathbb{R}^K \to \mathbb{R} : \boldsymbol{x} \to \max_{k=1,\ldots,K} x_k$ preserves comparisons in the convex increasing ordering. ■

## 10 Light traffic via the Reiman-Simon theory

We now shift attention to the light traffic regime. This refers to the limiting situation where the system traffic intensity approaches zero. Throughout this section we assume that the customer arrival process is *Poisson*. In that case the Reiman–Simon theory [12, 13] applies, and enables us to calculate derivatives of system quantities of interest with respect to the intensity of the Poisson arrival process, when this intensity tends to zero. Here, the quantity of interest is $\mathbf{P}_\lambda\,[v_\infty > x]$ ($x \geq 0$), the complementary cumulative distribution function of the stationary end-to-end delay. Our objective is to compute its derivatives of order zero and one, with a view towards establishing asymptotic optimality of Round–Robin routing in light traffic. To that end we next highlight the key points of the Reiman–Simon method, as it applies in our context.

As in Section 4, we start by introducing bi–infinite counterparts to the sequences of $\mathbb{R}_+$–valued rvs representing customer arrival epochs and their service durations. That

is, we consider the sample space $(\Omega, \mathcal{F})$ where $\Omega$ is the set of all finite and infinite sequences $\{(t_n, \sigma_n),\ n = 0, \pm 1, \pm 2, \ldots\}$ with $\sigma_n$ in $\mathbb{R}_+$ and $\ldots < t_{-1} < 0 = t_0 < t_1 < \ldots$. For $\lambda > 0$, we introduce a probability measure $\mathbf{P}_\lambda$ on $(\Omega, \mathcal{F})$ such that under $\mathbf{P}_\lambda$, the bi–infinite sequence of interrarival times $\{\tau_n,\ n = 0, \pm 1, \pm 2, \ldots\}$, given by $\tau_n = t_n - t_{n-1}$, is independent, exponentially distributed with parameter $\lambda$, and the marks $\{\sigma_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ are i.i.d. with common distribution $G$ which are independent of $\{\tau_n,\ n = 0, \pm 1, \pm 2, \ldots\}$. We also introduce the bi–infinite sequences of routing vectors $\{\boldsymbol{u}_n,\ n = 0, \pm 1, \pm 2, \ldots\}$, workload vectors $\{\boldsymbol{w}_n,\ n = 0, \pm 1, \pm 2, \ldots\}$, and end-to-end delays $\{v_n,\ n = 0, \pm 1, \pm 2, \ldots\}$. In this setup, the system has been operating from time $t = -\infty$, i.e., for each $n = 0, \pm 1, \pm 2, \ldots$,

$$w_{n+1}^k = [w_n^k + u_n^k \sigma_n - \tau_{n+1}]^+, \quad k = 1, \ldots, K \tag{64}$$

and

$$v_n := \max_{k=1,\ldots,K} \left( w_n^k + u_n^k \sigma_n \right). \tag{65}$$

Under the stability condition $\lambda \mathbf{E}[\sigma] < K$, convergence to the stationary rv $v_\infty$ has taken place by time $t = 0$, i.e.,

$$v_0 =_{st} v_\infty. \tag{66}$$

Let the generic system performance metric $\phi(\lambda)$ be expressed as

$$\phi(\lambda) = \int \Phi \, d\mathbf{P}_\lambda \tag{67}$$

for a suitably chosen rv $\Phi : \Omega \to \mathbb{R}$. For example, $\Phi(\omega)$ can be chosen as the system time $v_0$ of the tagged customer arriving at time $t = 0$, corresponding to a sample path $\omega$ in $\Omega$, in which case, from (66) and (67), the performance metric $\phi(\lambda)$ is the expected value $\mathbf{E}_\lambda[v_\infty]$ of the stationary end-to-end delay $v_\infty$. Application of the Reiman–Simon method entails conditioning on the number of arriving customers and their corresponding service times. To do so, we denote by $\emptyset$ the event where no customer arrives on the entire real line $(-\infty, \infty)$, *except* for a tagged customer who arrives at time $t = 0$ with service time $\sigma_0$. For each $t$ in $\mathbb{R}$, let $\{t\}$ denote the event where in addition to the tagged customer, there is exactly one more arrival occuring at time $t$; its service time is denoted by $\sigma$.

Next, we associate with $\Phi$ several auxiliary functions, namely the expected values of $\Phi$, conditionally on the arrival events $\emptyset$ and $\{t\}$, given by

$$\widehat{\Phi}(\emptyset) := \mathbf{E}_\lambda[\Phi \mid \emptyset] \quad \text{and} \quad \widehat{\Phi}(\{t\}) := \mathbf{E}_\lambda[\Phi \mid \{t\}], \quad t \in \mathbb{R}. \tag{68}$$

These quantities do *not* depend on $\lambda$.

The following result is essentially a version of Theorem 2 given in [13, p. 30]. It provides formulas for the derivatives of order zero and one of $\phi(\lambda)$ at $\lambda = 0+$, by considering scenarios where, in addition to the tagged customer, at most one more customer ever joins the system.

**Proposition 10.1** *If there exists $\theta^\star > 0$ such that $\mathbf{E}\left[e^{\theta\sigma}\right] < \infty$ for $\theta < \theta^\star$, then*

$$\lim_{\lambda \downarrow 0} \phi(\lambda) = \widehat{\Phi}(\emptyset) \tag{69}$$

*and*

$$\frac{d}{d\lambda}\, \phi(0+) := \lim_{\lambda \downarrow 0} \frac{d}{d\lambda}\, \phi(\lambda) = \int_{\mathbf{R}} \left(\widehat{\Phi}(\{t\}) - \widehat{\Phi}(\emptyset)\right)\, dt. \tag{70}$$

# 11 Light traffic optimality of Round–Robin routing

We rely on Proposition 10.1 to calculate light traffic derivatives of the distribution of the end-to-end delay. Using these derivatives we write a Taylor expansion of the end-to-end delay distribution around $\lambda = 0$ given in the next proposition. This expansion is valid for all admissible routing policies (prescribed by Assumption (A)), i.e., it is not limited to Round–Robin or Bernoulli routing. It depends critically on the event that the routing policy assigns two successive customer to the same queue. Throughout, we write $\bar{G}(x) := 1 - G(x)$ $(x \geq 0)$ for the complementary cumulative service time distribution.

**Proposition 11.1** *Assume Poisson arrivals of rate $\lambda$ and finite exponential moments for the service time distribution as in Proposition 10.1. If $G(0) = 0$, then the distribution of the end-to-end delay under any admissible routing policy is given by*

$$\begin{aligned} \mathbf{P}_{\lambda,\alpha}\left[v_\infty > x\right] &= \bar{G}(x) + \alpha\lambda \left(\mathbf{E}\left[\sigma\right] G(x) - \int_0^x G(x-y)\bar{G}(y)\, dy\right) \\ &\quad +(1-\alpha)\lambda G(x) \int_x^\infty \bar{G}(y)\, dy + o(\lambda), \quad x \geq 0, \end{aligned} \tag{71}$$

*where*

$$\alpha := \mathbf{P}\left[\nu_{-1} = \nu_0\right] \tag{72}$$

*denotes the probability that the routing policy assigns two consecutive arrivals to the same queue.*

The proof of Proposition 11.1 is given in Appendix A.4. The expansions under Round–Robin and Bernoulli routing can be obtained as easy special cases:

**Corollary 11.1** *In the setup of Proposition 11.1, under any admissible routing policy it holds that*

$$\mathbf{P}_\lambda\left[v_\infty^B > x\right] = \mathbf{P}_{\lambda,1/K}\left[v_\infty > x\right] \quad \textit{and} \quad \mathbf{P}_\lambda\left[v_\infty^R > x\right] = \mathbf{P}_{\lambda,0}\left[v_\infty > x\right]. \tag{73}$$

**Proof.** Conclusion (73) follows immediately from (71) by noting that $\mathbf{P}\left[\nu_{-1}^R = \nu_0^R\right] = 0$ and $\mathbf{P}\left[\nu_{-1}^B = \nu_0^B\right] = 1/K$. ■

Observe that (71) can be explained by considering the light traffic situation where at most one more customer and the tagged customer join the system. These two customers are assigned to different queues with probability $1-\alpha$, in which case, from the tagged customer's perspective, any routing policy is tantamount to Round–Robin routing. On the other hand, with probability $\alpha$, both customers are assigned to the same queue, in which case the system behaves like an $M|G|1$ queue. Thus, up to first order in $\lambda$, the end-to-end delay performance under any routing policy is a mixture of these two components. Such an interpretation is confirmed upon comparing the corresponding terms in (71) with the known expansion for the $M|G|1$ queue [9, p. 201] and the expression in (73) for Round–Robin routing. The details are given in Appendix A.4.

Proposition 11.1 leads to the conclusion that in light traffic the Round–Robin policy is an optimal routing policy, although this optimality is shared by a number of policies, namely all the admissible policies for which $\alpha = 0$.

**Corollary 11.2** *In the setup of Proposition 11.1, it holds that*

$$\lim_{\lambda\downarrow 0}\frac{1}{\lambda}\left(\mathbf{P}_{\lambda,\alpha}\left[v_\infty > x\right] - \mathbf{P}_{\lambda,0}\left[v_\infty > x\right]\right) \geq 0, \quad x \geq 0. \tag{74}$$

*In other words, Round - Robin is optimal among all admissible routing policies in the sense that $v_\infty^R \leq_{st} v_\infty$ asymptotically in light traffic.*

**Proof.** Fix $x \geq 0$. Inequality (74) follows from (71) as we get

$$\begin{aligned}
&\lim_{\lambda\downarrow 0}\frac{1}{\lambda}\left(\mathbf{P}_{\lambda,\alpha}\left[v_\infty > x\right] - \mathbf{P}_{\lambda,0}\left[v_\infty > x\right]\right)\\
= \;& \alpha\left(\mathbf{E}\left[\sigma\right]G(x) - G(x)\int_x^\infty \bar{G}(y)\,dy - \int_0^x G(x-y)\bar{G}(y)\,dy\right)\\
\geq \;& \alpha\left(\mathbf{E}\left[\sigma\right]G(x) - G(x)\int_x^\infty \bar{G}(y)\,dy - G(x)\int_0^x \bar{G}(y)\,dy\right) = 0
\end{aligned}$$

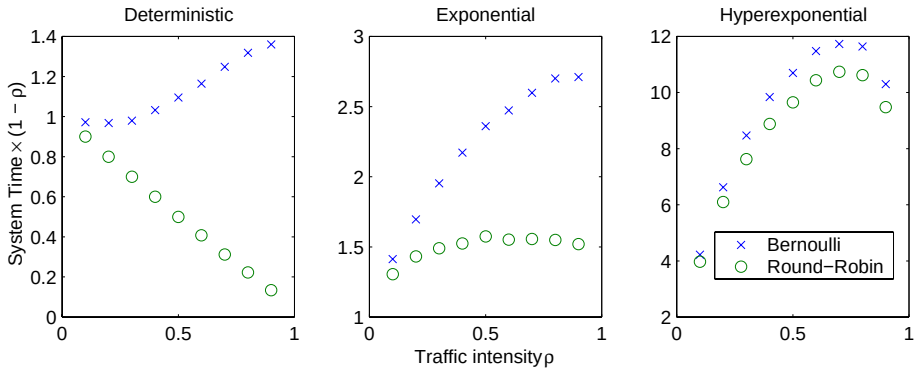


Figure 1: Expected end-to-end delay; Bernoulli vs Round–Robin routing; $K = 10$

by the monotonicity of $G$. The ensuing stochastic comparison is a consequence of identification (73). ■

## 12 Simulation results

To illustrate the asymptotic optimality of Round – Robin policy we present numerical examples comparing a system of parallel queues with Round–Robin routing to a system with Bernoulli routing. In all the examples shown below customers arrive to the parallel queueing system according to a Poisson process. We perform simulation experiments using three different distributions for the service time rv $\sigma$. In particular, we consider, in order of increasing variability, the deterministic distribution $D(x) = 1$ ($x \geq 1$), the exponential distribution $\mathcal{E}(x) = 1 - e^{-x}$ ($x \geq 0$), and the hyperexponential distribution (mixture of three exponentials) given by

$$H(x) = \sum_{i=1}^{3} p_i(1 - e^{-x/f_i}), \quad x \geq 0$$

with $p_1 = 0.3$, $p_2 = 0.6$, $p_3 = 0.1$, and $f_1 = 0.2$, $f_2 = 0.5$, $f_3 = 6.4$, yielding a squared coefficient of variation $c_H^2 = 7.516$. In all three cases the expected service time is equal to one.

In Figure 1 we vary the Poisson arrival rate $\lambda$ and plot the expected end-to-end delay, times $(1 - \rho)$, against the system utilization $\rho$, for a system of $K = 10$ parallel queues with resequencing. We see that the expected end-to-end delay under Round–Robin routing is smaller than its counterpart under Bernoulli routing: This conclusion holds for all

three service distributions and across all traffic intensities, not just in heavy or light traffic. Tables 1 and 2 show that such a comparison remains true when we vary the number of parallel queues from $K = 2$ to $K = 10$. We conjecture that, for the case of Poisson arrivals, a comparison in the increasing convex stochastic ordering between the end-to-end delays under Round–Robin and Bernoulli routing, holds true across all traffic intensities. Figure 1, together with Tables 1 and 2, provide evidence, though circumstantial, in support of this conjecture. It is also worthwhile to note, from Table 2, that in the case of deterministic service times the end-to-end delay under Round–Robin routing is a decreasing function of the number $K$ of parallel queues (while it is eventually increasing in the limit as $K$ grows unboundedly for all other distributions). This is due to the fact that, in the case of deterministic service times, customers experience no resequencing delay under Round–Robin routing. Furthermore, their Erlang interarrival times to each queue tend to deterministic of duration $\mathbf{E}[\sigma]/\rho$ as $K$ increases to infinity.

| | | Expected end-to-end delay $\mathbf{E}\left[v_\infty^B\right]$ | | | | |
|---|---|---|---|---|---|---|
| $\rho$ | Service d.f. | $K=2$ | $K=4$ | $K=6$ | $K=8$ | $K=10$ |
| 0.3 | $D$ | 1.26 | 1.30 | 1.33 | 1.37 | 1.40 |
| | $\mathcal{E}$ | 1.67 | 2.00 | 2.31 | 2.59 | 2.79 |
| | $H$ | 4.27 | 6.78 | 8.69 | 10.5 | 12.1 |
| 0.7 | $D$ | 2.58 | 3.14 | 3.52 | 3.91 | 4.16 |
| | $\mathcal{E}$ | 4.84 | 6.10 | 7.22 | 8.11 | 8.66 |
| | $H$ | 17.6 | 25.9 | 31.6 | 36.2 | 39.1 |

Table 1: End-to-end delay under Bernoulli routing.

# A Appendix

## A.1 A proof of Lemma 7.1

Fix $r = 1, 2, \ldots$ and $k, \ell = 1, \ldots, K$. We start with the obvious decomposition

$$\operatorname{cov}\left[\sum_{n=0}^{r-1} u_n^k \sigma_n, \sum_{m=0}^{r-1} u_m^\ell \sigma_m\right] = \sum_{n=0}^{r-1}\sum_{m=0}^{r-1} \operatorname{cov}\left[u_n^k \sigma_n, u_m^\ell \sigma_m\right]. \tag{A.1}$$

Fixing $n, m = 0, 1, \ldots$, we note under the independence in Assumption (A) that

$$\operatorname{cov}\left[u_n^k \sigma_n, u_m^\ell \sigma_m\right] \;=\; \mathbf{E}\left[u_n^k u_m^\ell \sigma_n \sigma_m\right] - \mathbf{E}\left[u_n^k \sigma_n\right] \mathbf{E}\left[u_m^\ell \sigma_m\right]$$

| | | Expected end-to-end delay $\mathbf{E}\left[v_\infty^R\right]$ | | | | |
|---|---|---|---|---|---|---|
| $\rho$ | Service d.f. | $K=2$ | $K=4$ | $K=6$ | $K=8$ | $K=10$ |
| 0.3 | $D$ | 1.06 | 1.01 | 1.00 | 1.00 | 1.00 |
| | $\mathcal{E}$ | 1.45 | 1.64 | 1.83 | 2.00 | 2.13 |
| | $H$ | 4.05 | 6.29 | 8.12 | 9.66 | 10.9 |
| 0.7 | $D$ | 1.51 | 1.19 | 1.10 | 1.06 | 1.04 |
| | $\mathcal{E}$ | 3.47 | 4.01 | 4.43 | 4.85 | 5.19 |
| | $H$ | 16.7 | 24.4 | 29.1 | 32.8 | 35.8 |

Table 2: End-to-end delay under Round–Robin routing.

$$\begin{aligned} &= \mathbf{E}\left[u_n^k u_m^\ell\right]\mathbf{E}\left[\sigma_n\sigma_m\right] - \mathbf{E}\left[u_n^k\right]\mathbf{E}\left[u_m^\ell\right]\mathbf{E}\left[\sigma_n\right]\mathbf{E}\left[\sigma_m\right] \\ &= \mathbf{E}\left[u_n^k u_m^\ell\right]\left(\delta(n,m)\mathbf{E}\left[\sigma^2\right] + (1-\delta(n,m))\mathbf{E}\left[\sigma\right]^2\right) \\ &\quad -\mathbf{E}\left[u_n^k\right]\mathbf{E}\left[u_m^\ell\right]\mathbf{E}\left[\sigma\right]^2 \\ &= \delta(n,m)\mathbf{E}\left[u_n^k u_m^\ell\right]\mathrm{var}\left[\sigma\right] + \mathrm{cov}\left[u_n^k, u_m^\ell\right]\cdot\mathbf{E}\left[\sigma\right]^2. \end{aligned}$$

Moreover, it is also the case that

$$\begin{aligned} \delta(n,m)\mathbf{E}\left[u_n^k u_m^\ell\right] &= \delta(n,m)\mathbf{E}\left[u_n^k u_n^\ell\right] \\ &= \delta(n,m)\delta(k,\ell)\mathbf{E}\left[u_n^k\right] \\ &= \delta(n,m)\delta(k,\ell)\frac{1}{K}, \end{aligned} \tag{A.2}$$

so that

$$\mathrm{cov}\left[u_n^k\sigma_n, u_m^\ell\sigma_m\right] = \delta(n,m)\delta(k,\ell)\frac{\mathrm{var}\left[\sigma\right]}{K} + \mathrm{cov}\left[u_n^k, u_m^\ell\right]\cdot\mathbf{E}\left[\sigma\right]^2. \tag{A.3}$$

Substituting (A.3) into (A.1) yields

$$\begin{aligned} &\mathrm{cov}\left[\sum_{n=0}^{r-1} u_n^k\sigma_n, \sum_{m=0}^{r-1} u_m^\ell\sigma_m\right] \\ &= \delta(k,\ell)r\cdot\frac{\mathrm{var}\left[\sigma\right]}{K} + \mathbf{E}\left[\sigma\right]^2\cdot\sum_{n=0}^{r-1}\sum_{m=0}^{r-1}\mathrm{cov}\left[u_n^k, u_m^\ell\right] \\ &= \delta(k,\ell)r\cdot\frac{\mathrm{var}\left[\sigma\right]}{K} + \mathbf{E}\left[\sigma\right]^2\cdot\mathrm{cov}\left[\sum_{n=0}^{r-1} u_n^k, \sum_{m=0}^{r-1} u_m^\ell\right]. \end{aligned} \tag{A.4}$$

It is now plain that the $K \times K$ covariance matrix $\Gamma$ at (37) exists if and only if the $K \times K$ covariance matrix $\widetilde{\Gamma}$ at (40) exists, in which case the relation (41) is immediate. This concludes the proof of Lemma 7.1. ■

## A.2 A proof of Proposition 8.1

**Bernoulli routing:** Under Bernoulli routing, the FCLT (36) is essentially the $K$–dimensional version of Donsker's Theorem [3, Thm. 16.1, p. 137] when applied to the i.i.d. rvs $\{\boldsymbol{u}_n\sigma_n,\ n = 0, 1, \ldots\}$. In order to determine the form of $\widetilde{\Gamma}^B$, fix $k, \ell = 1, \ldots, K$. For arbitrary $n, m = 0, 1, \ldots$, by independence we note that

$$\begin{aligned} \mathrm{cov}\left[u_n^{B,k}, u_m^{B,\ell}\right] &= \delta(n,m)\mathrm{cov}\left[u_n^{B,k}, u_n^{B,\ell}\right] \\ &= \delta(n,m)\delta(k,\ell)\cdot\frac{1}{K}\left(1-\frac{1}{K}\right). \end{aligned} \tag{A.5}$$

Expression (42) readily follows upon using (A.5) in the right-hand side of relation[2]

$$\frac{1}{r}\mathrm{cov}\left[\sum_{n=0}^{r-1} u_n^{B,k}, \sum_{m=0}^{r-1} u_m^{B,k}\right] = \frac{1}{r}\sum_{n=0}^{r-1}\sum_{m=0}^{r-1}\mathrm{cov}\left[u_n^{B,k}, u_m^{B,\ell}\right], \quad r = 1, 2, \ldots$$

and then letting $r$ go to infinity.

**Round–Robin routing:** Fix $r = 1, 2, \ldots$ and $t > 0$. We begin by rewriting (35) as

$$\boldsymbol{U}^{R,r}(t) = \boldsymbol{Z}^r(t) + \frac{1}{\sqrt{r}}\sum_{n=0}^{[rt]-1}\left(\boldsymbol{u}_n^R - \frac{1}{K}\mathbf{1}\right)\cdot\mathbf{E}\left[\sigma\right] \tag{A.6}$$

where the $\mathbb{R}^K$-valued process $\{\boldsymbol{Z}^r(t),\ t \geq 0\}$ is defined componentwise by

$$Z^{r,k}(t) := \frac{1}{\sqrt{r}}\sum_{n=0}^{[rt]-1}\mathbf{1}\left[\nu_n^R = k\right](\sigma_n - \mathbf{E}\left[\sigma\right]), \quad t \geq 0,\ k = 1, 2, \ldots, K \tag{A.7}$$

with routing rvs $\{\nu_n^R,\ n = 0, 1, \ldots\}$ given by (6). The arguments proceed along a number of steps.

**Step 1:** Fix $r = 1, 2, \ldots$ and $t > 0$. For each $k = 1, \ldots, K$, we note that

$$\frac{[rt]}{K} - 1 < \sum_{n=0}^{[rt]-1} u_n^{R,k} \leq \frac{[rt]}{K} + 1, \tag{A.8}$$

[2]This relation was used already on the way to (A.4).

whence

$$-\frac{1}{\sqrt{r}} < \frac{1}{\sqrt{r}} \sum_{n=0}^{[rt]-1} \left( u_n^{R,k} - \frac{1}{K} \right) \leq -\frac{1}{\sqrt{r}}. \tag{A.9}$$

It is now immediate that for each $T > 0$,

$$\lim_{r\to\infty} \sup_{0<t\leq T} \max_{k=1,\ldots,K} \left| \frac{1}{\sqrt{r}} \sum_{n=0}^{[rt]-1} \left( u_n^{R,k} - \frac{1}{K} \right) \right| = 0. \tag{A.10}$$

Consequently, by Theorem 4.1 in [3, p. 25], the desired convergence (36), with covariance matrix as in (43), will hold if we show instead the convergence

$$\{\boldsymbol{Z}^r(t),\ t \geq 0\} \Longrightarrow_r \{(\Gamma^R)^{1/2}\boldsymbol{A}(t),\ t \geq 0\} \tag{A.11}$$

where $\{\boldsymbol{A}(t),\ t \geq 0\}$ is a $K$–dimensional standard Brownian motion, and $\Gamma^R$ is the $K \times K$ covariance matrix given by

$$\Gamma_{k\ell}^R = \delta(k,\ell)\frac{\text{var}\,[\sigma]}{K}, \quad k,\ell = 1,\ldots,K, \tag{A.12}$$

where we have taken into account (41).

**Step 2:** Fix $r = 1, 2, \ldots$ and $k, \ell = 1, \ldots, K$. We have

$$\begin{aligned} & \frac{1}{r}\text{cov}\left[\sum_{n=0}^{r-1} u_n^{R,k}, \sum_{m=0}^{r-1} u_m^{R,\ell}\right] \\ = \quad & \mathbf{E}\left[\sqrt{r}\left(\frac{1}{r}\sum_{n=0}^{r-1} u_n^{R,k} - \frac{1}{K}\right) \cdot \sqrt{r}\left(\frac{1}{r}\sum_{m=0}^{r-1} u_m^{R,\ell} - \frac{1}{K}\right)\right] \end{aligned} \tag{A.13}$$

upon using the stationarity assumption on the assignment sequence (with (5)). Let $r$ go to infinity in (A.13) and observe from (A.9) (with $t = 1$) that

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{1}{r}\sum_{n=0}^{r-1} u_n^{R,k} - \frac{1}{K}\right) = 0, \quad k = 1,\ldots,K. \tag{A.14}$$

We readily conclude from definition (40) that $\widetilde{\Gamma}_{k\ell}^R = 0$ by the Bounded Convergence Theorem (with the help of bounds (A.8) with $t = 1$). This establishes (A.12) via (41).

**Step 3:** We now turn to showing (A.11). A moment of reflection should convince the reader that for each $k = 1, 2, \ldots, K$, *conditionally* on $\nu^\star = \nu_0^R = k$, the $K$ processes $\{Z^{r,1}(t),\ t \geq 0\}$, …, $\{Z^{r,K}(t),\ t \geq 0\}$ are mutually independent for each $r = 1, 2, \ldots$.

Appealing to the cyclical nature of the Round-Robin policy, we readily conclude to the convergence

$$\{\boldsymbol{Z}^r(t),\ t \geq 0\} \Longrightarrow_r \left\{ \sqrt{\frac{\mathrm{var}\,[\sigma]}{K}}\boldsymbol{A}(t),\ t \geq 0 \right\}, \tag{A.15}$$

*conditionally* on $\nu_0^R = k$. This is an easy consequence of Donsker's Theorem [3, Thm. 16.1, p. 137]; details are left to the interested reader. As the limit in (A.15) does not depend on $k$, it follows that this convergence holds *unconditionally* as well! This last convergence is essentially (A.11) as we recall the expression (A.12) for $\Gamma^R$. ■

## A.3 A proof of Proposition 9.2

Clearly, $\{\boldsymbol{S}(t),\ t \geq 0\}$ is a zero drift $K$–dimensional Brownian motions, so that $\boldsymbol{S}(t_1, \ldots, t_n)$ are $nK$–dimensional zero mean Gaussian rvs. Their $nK \times nK$ covariance matrix has a block structure with the $(i, j)$ block being the $K \times K$ matrix given by

$$\min(t_i, t_j)\Sigma, \quad i, j = 1, \ldots, n, \tag{A.16}$$

with the $K \times K$ matrix $\Sigma$ as in Proposition 8.2. With Proposition 9.1 in mind, we see that the extremal property (58) of Round–Robin routing will be established if we show that the matrix difference

$$C(t_1, \ldots, t_n) := \mathrm{cov}[\boldsymbol{S}(t_1, \ldots, t_n)] - \mathrm{cov}[\boldsymbol{S}^R(t_1, \ldots, t_n)]$$

is positive semi-definite. To this end, write an arbitrary element $\boldsymbol{v}$ in $\mathbb{R}^{nK}$ in block form as $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ with $\boldsymbol{v}_i$ a vector of $\mathbb{R}^K$ for each $i = 1, \ldots, n$. We need to show that

$$\boldsymbol{v}C(t_1, \ldots, t_n)\boldsymbol{v}^T = \sum_{i=1}^{n}\sum_{j=1}^{n} \min(t_i, t_j) \cdot \boldsymbol{v}_i(\Sigma - \Sigma^R)\boldsymbol{v}_j^T \geq 0. \tag{A.17}$$

Expression (47) immediately yields

$$\Sigma - \Sigma^R = \left(\widetilde{\Gamma} - \widetilde{\Gamma}^R\right) \cdot \mathbf{E}\,[\sigma]^2 = \widetilde{\Gamma} \cdot \mathbf{E}\,[\sigma]^2$$

with the help of (43). By virtue of its definition (40), the matrix $\widetilde{\Gamma}$ is itself a covariance matrix, whence is positive semi-definite. Thus, for $n = 1$, we have that (A.17) holds as it is equivalent to $\widetilde{\Gamma}$ being positive semi-definite.

By standard results from linear algebra [17, p. 339], there exists a $K \times K$ orthonormal matrix $P$ such that the matrix $\Lambda = P\widetilde{\Gamma}P^T$ is diagonal with non-negative diagonal elements

$\lambda_1, \ldots, \lambda_K$. Making the change of variables $\boldsymbol{w}_i = \boldsymbol{v}_i P^T = (w_i^1, \ldots, w_i^K)$, or equivalently, $\boldsymbol{v}_i = \boldsymbol{w}_i P$, for each $i = 1, \ldots, n$, we find

$$\begin{aligned}
\frac{1}{\mathbf{E}\left[\sigma\right]^2}\boldsymbol{v}C(t_1, \ldots, t_n)\boldsymbol{v}^T &= \sum_{i=1}^{n}\sum_{j=1}^{n}\min(t_i, t_j)\cdot \boldsymbol{w}_i P\widetilde{\Gamma}P^T\boldsymbol{w}_j^T \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\min(t_i, t_j)\cdot \boldsymbol{w}_i \Lambda \boldsymbol{w}_j^T \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\min(t_i, t_j)\cdot \sum_{k=1}^{K}\lambda_k w_i^k w_j^k \\
&= \sum_{k=1}^{K}\lambda_k\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\min(t_i, t_j) w_i^k w_j^k\right). \qquad \text{(A.18)}
\end{aligned}$$

For each $k = 1, \ldots, K$, we introduce the element $\boldsymbol{w}_k^\star = (w_1^k, \ldots, w_n^k)$ of $\mathbb{R}^n$ to write

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\min(t_i, t_j) w_i^k w_j^k = \boldsymbol{w}_k^\star \mathrm{cov}[B(t_1, \ldots, t_n)](\boldsymbol{w}_k^\star)^T \qquad \text{(A.19)}$$

where $\mathrm{cov}[B(t_1, \ldots, t_n)]$ denotes the covariance matrix of the vector $B(t_1, \ldots, t_n) := (B(t_1), \ldots, B(t_n))$ with $\{B(t),\ t \geq 0\}$ a standard one-dimensional Brownian motion. A covariance matrix being positive semi-definite, each of the terms at (A.18) is non-negative, whence (A.17) holds by virtue of the fact that $\lambda_k \geq 0$ for all $k = 1, \ldots, K$. ■

## A.4 A proof of Proposition 11.1

Fix $x \geq 0$ and take $\Phi = \mathbf{1}\left[v_0 > x\right]$. The following holds irrespective of the customer assignments used: First, on the event $\emptyset$, we have $v_0 = \sigma_0$, so that

$$\widehat{\Phi}(\emptyset) = \mathbf{P}\left[\sigma > x\right] = \bar{G}(x). \qquad \text{(A.20)}$$

Next, with $t > 0$, on the event $\{t\}$ it is also the case that $v_0 = \sigma_0$, whence $\widehat{\Phi}(\{t\}) = \widehat{\Phi}(\emptyset)$, and we conclude from (70) that

$$\frac{d}{d\lambda}\,\phi(0+) = \int_{-\infty}^{0}\left(\widehat{\Phi}(\{t\}) - \widehat{\Phi}(\emptyset)\right)\,dt. \qquad \text{(A.21)}$$

Finally, fix $t < 0$ and observe that

$$v_0 = \max_{k=1,\ldots,K}\left([\delta(k, \nu_{-1})\sigma + t]^+ + \delta(k, \nu_0)\sigma_0\right). \qquad \text{(A.22)}$$

The remainder of the proof consists in evaluating $\widehat{\Phi}(\{t\})$ ($t < 0$) through (A.22). On the event $\nu_{-1} \neq \nu_0$ (this is tantamount to Round–Robin routing), (A.22) becomes

$$v_0 = \max\left([\sigma + t]^+, \sigma_0\right), \tag{A.23}$$

while on the event $\nu_{-1} = \nu_0$ (which corresponds to the $M|G|1$ queue), it holds that

$$v_0 = [\sigma + t]^+ + \sigma_0. \tag{A.24}$$

Using this information, by the definition (72) for $\alpha$ and the admissibility of the routing poilicy, we get

$$\widehat{\Phi}(\{t\}) = \alpha \mathbf{P}\left[[\sigma + t]^+ + \sigma_0 > x\right] + (1-\alpha)\mathbf{P}\left[\max\left([\sigma + t]^+, \sigma_0\right) > x\right] \tag{A.25}$$

with

$$\begin{aligned}
\mathbf{P}\left[\max\left([\sigma + t]^+, \sigma_0\right) > x\right] &= 1 - \mathbf{P}\left[\sigma_0 \leq x\right]\mathbf{P}\left[[\sigma + t]^+ \leq x\right] \\
&= 1 - \mathbf{P}\left[\sigma_0 \leq x\right]\left(1 - \mathbf{P}\left[\sigma + t > x\right]\right) \\
&= 1 - G(x)\left(1 - \bar{G}(x - t)\right) \\
&= \bar{G}(x) + G(x)\bar{G}(x - t).
\end{aligned} \tag{A.26}$$

Next, recalling (A.20), we find

$$\begin{aligned}
&\widehat{\Phi}(\{t\}) - \widehat{\Phi}(\emptyset) \\
= \;& \alpha \mathbf{P}\left[[\sigma + t]^+ + \sigma_0 > x\right] + (1-\alpha)\left(\bar{G}(x) + G(x)\bar{G}(x - t)\right) - \bar{G}(x) \\
= \;& \alpha\left(\mathbf{P}\left[[\sigma + t]^+ + \sigma_0 > x\right] - \bar{G}(x)\right) + (1-\alpha)G(x)\bar{G}(x - t)
\end{aligned} \tag{A.27}$$

with the help of the calculations leading to (A.26). We further write

$$\begin{aligned}
\mathbf{P}\left[[\sigma + t]^+ + \sigma_0 > x\right] - \bar{G}(x) &= \mathbf{P}\left[\sigma_0 \leq x < [\sigma + t]^+ + \sigma_0\right] \\
&= \mathbf{E}\left[\mathbf{1}\left[\sigma_0 \leq x\right]\bar{G}(x - \sigma_0 - t)\right],
\end{aligned} \tag{A.28}$$

and reporting (A.27) and (A.28) into (A.21), we obtain

$$\begin{aligned}
&\frac{d}{d\lambda}\,\phi(0+) \\
= \;& (1-\alpha)G(x)\int_x^\infty \bar{G}(y)\,dy + \alpha\int_0^\infty \mathbf{E}\left[\mathbf{1}\left[\sigma_0 \leq x\right]\bar{G}(x - \sigma_0 + y)\right]\,dy.
\end{aligned}$$

Expansion (71) now follows once we observe that

$$\begin{aligned}
& \int_0^\infty \mathbf{E}\left[\mathbf{1}\left[\sigma_0 \le x\right] \bar{G}(x-\sigma_0+y)\right]\, dy \\
= & \int_0^\infty dy \int_0^x dG(z) \bar{G}(x-z+y) \\
= & \int_0^x dG(z) \int_{x-z}^\infty \bar{G}(y)\, dy \\
= & \int_0^x dG(z) \int_0^\infty \bar{G}(y)\, dy - \int_0^x dG(z) \int_0^{x-z} \bar{G}(y)\, dy \\
= & \mathbf{E}\left[\sigma\right] G(x) - \int_0^x \bar{G}(y)\, dy \int_0^{x-y} dG(z)
\end{aligned}$$

since $G(0) = 0$. ■

## Acknowledgments

The authors thank Alfred Müller for making a preprint of [11] available. They also thank the anonymous reviewer for his/her careful reading of the manuscript, and for comments that lead to sharper results regarding the asymptotic optimality of the Round–Robin policy.

## References

[1] S. Asmussen. *Applied Probability and Queues*. Wiley, Chichester, West Sussex, UK, 1987.

[2] F. Baccelli and A. M. Makowski. Queueing models for systems with synchronization constraints. In *Proceedings of the IEEE 77, Special Issue on Discrete Event Systems*, pages 138–161, New York, January 1989.

[3] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York (NY), 1968.

[4] I. I. Gikhman and A. V. Skorokhod. *An Introduction to the Theory of Random Processes*. Saunders Company, 1969.

[5] N. Gogate and S. S. Panwar. Assigning customers to two parallel servers with resequencing. *IEEE Communications Letters*, 3(4):119–122, April 1999.

[6] J. M. Harrison, *Brownian Motion and Stochastic Flow Systems*. John Wiley & Sons, New York (NY), 1985.

[7] I. Iliadis and L.Y.-C. Lien. Resequencing delay for a queueing system with two heterogeneous servers under a threshold type scheduling. *IEEE Transactions on Communications*, 36(6):692–702, June 1988.

[8] A. Jean-Marie and L. Gün. Parallel queues with resequencing. *Journal of the ACM*, 40(5):1188–1208, November 1993.

[9] L. Kleinrock. *Queueing Systems, Volume I: Theory*. John Wiley and Sons, New York (NY), 1975.

[10] R. M. Loynes. The stability of a queue with non–independent inter–arrival and service times. *Proceedings of the Cambridge Philosophical Society*, 58:497–520, 1962.

[11] A. Müller. Stochastic ordering of multivariate normal distributions. *Annals of the Institute of Statistical Mathematics*, 53:567-575, 2001.

[12] M. I. Reiman and B. Simon. Light traffic limits of sojourn time distributions in Markovian queueing networks. *Stochastic Models*, 4:191–233, 1988.

[13] M. I. Reiman and B. Simon. Open queueing systems in light traffic. *Mathematics of Operations Research*, 14:26–59, 1989.

[14] S. Ross. *Stochastic Processes*. John Wiley & Sons, New York (NY), 1984.

[15] M. Shaked and J. G. Shanthikumar. *Stochastic Orders and Their Applications*. Academic Press, New York (NY) 1994.

[16] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. English Translation (D.J. Daley, Editor). J. Wiley & Sons, New York (NY), 1984.

[17] G. Strang. *Linear Algebra and Its Applications*. Harcourt Brace Jovanovich, Third Edition, San Diego (CA), 1988.

[18] S. Varma. Optimal allocation of customers in a two server queue with resequencing. *IEEE Transactions on Automatic Control*, 36(11):1288–1293, November 1991.

[19] W. Whitt. Heavy traffic limit theorems in queues: A survey. In A. B. Clarke, editor, *Mathematical methods in queueing theory*, Lecture Notes in Economics and Mathematical Systems 98, pages 307–350, Berlin, 1974. Springer–Verlag.

[20] W. Whitt. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, 5:67–85, 1980.

[21] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Applications to Queues*. Springer-Verlag, New York (NY), 2002.